BU - 1138 - M

Draft:

Sampling Based Methods for the Estimation of DNA Sequence Accuracy

Gary Churchill George Casella

January 4, 1993

Abstract

We present a model for random errors which occur in DNA sequence data. The model is defined in terms of three parameters, one for each of the possible error types, substitution, insertion or deletion. A resampling algorithm is described which can be used to simultaneously estimate the error rate parameters and to restore the DNA sequence. Parameter estimates are summarized as a posterior density. The restored DNA sequence can be summarized as a modal sequence or as a

posterior credible region which takes the form of a cylinder set in sequence space. The methods are applied to a set of sequenced fragments from *Escherichia coli*. Limitations and possible generalizations of the algorithm are discussed.

1 Introduction

Since the advent of rapid DNA sequencing technologies in the late 1970's, the rate of accumulation of DNA data has undergone an exponential growth. Although the existing data are not of uniform accuracy, this is generally not apparent from the database entries (see Waterman, 1990). There has been considerable recent debate regarding the acceptable levels of accuracy and how to estimate accuracy (Roberts, 1990). We will present a model for DNA sequencing errors and a resampling algorithm for simultaneous estimation of error rate parameters and restoration of the DNA sequence. The restored DNA sequence can be represented as a posterior credible region which takes the form of a cylinder set.

Large DNA sequences (up to 10^6 bases) are typically determined by assembling an overlapping set of smaller fragments (200–300 bases) each of which can be sequenced in a single experiment. Errors occur at the stage of determining the fragment sequences. The types of errors that may occur include

substitution of one DNA base for another, insertion of a base in the fragment which is not present in the true sequence and deletion of a base from the fragment. Many of these errors can be recognized as ambiguities in the overlapping portions of the assembled fragment set. However it is possible that some errors will remain undetected and appear in the final sequence.

The DNA molecule whose sequence is to be determined will be referred to as the clone sequence and is denoted by $\mathbf{s} = s_1, \ldots, s_n$. The individual bases in the clone sequence are elements of the alphabet $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ and its length n is an unknown finite integer. Thus $\mathbf{s}$ is an element of the space $\mathcal{S} = \bigcup_{k=1}^{\infty} \mathcal{A}^k$, where $\mathcal{A}^k$ are the sets of k–tuples on the alphabet $\mathcal{A}$. This space will be referred to as sequence space. The observable data consist of fragment sequences $\mathbf{f}_j = f_{1j}, \ldots, f_{n_j j}$, $j = 1, \ldots, m$ which are elements of sequence space with known lengths n_j. The fragment sequences are assumed to be assembled into a matrix $\mathbf{X}$ with elements x_{ij} as described below. The error rate parameters, denoted $\theta = (p_I, p_D, p_S)$, are the probabilities of insertion, deletion and substitution when a base in the clone sequence is copied to generate a base in a fragment sequence.

Our primary goal is to restore the clone sequence $\mathbf{s}$ given the assembled fragments matrix $\mathbf{X}$. A secondary goal is to estimate θ. If the error rates in fragment sequences are known, a probability distribution $\Pr(\mathbf{s}|\mathbf{X}, \theta)$ can be

defined over the space of sequences. Conversely, if the true sequence is known, a posterior distribution $\Pr(\theta|\mathbf{s}, \mathbf{X})$ for the error rates can be computed. By iteratively resampling from these conditional distributions we can obtain a sample from the marginal posterior $\Pr(s|x)$. Restoration of the clone sequence based on this sample accounts for uncertainty in the estimation of θ. This approach generalizes previous work (Churchill and Waterman, 1991) in which an EM algorithm is implemented to estimate θ and the sequence restoration is made conditional upon the MLE $\hat{\theta}$.

2 The Model

Let s_i denote the DNA base at a position in the clone sequence indexed by $i = 1, \ldots, n$. Each base is an element from the set $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. Immediately to the right of each base is an imaginary link which connects the DNA bases. We will refer to the links in the clone sequence as s–links. The concept of links is used to define the insertion–deletion error processes and was introduced by Thorne et al. (1991) in the context of a model for sequence evolution.

The individual DNA sequence fragments are generated by an error prone copying process on $\mathbf{s}$. Fragment sequence $\mathbf{f}_j$ (for $j = 1, \ldots, m$) is generated by the following process: starting at a given position $k = k(j)$ in the clone sequence,

1. The base s_k is copied. The copying has three possible outcomes.

   (a) A substitution occurs with probability p_S. The fragment element f_{ij} takes a value from the set $\mathcal{A} \setminus s_k$ with equally likely probabilities $1/3$ and the index i is incremented.

   (b) A deletion occurs with probability p_D. The fragment element is assigned no value and the s–link is also deleted.

   (c) A true copy occurs with probability $1 - p_S - p_D$. The fragment element f_{ij} takes the value s_k and the index i is incremented.

2. Given that the outcome of step 1 is not a deletion, the s–link is copied. There are two possible outcomes:

   (a) An insertion occurs with probability p_I. Multiple bases may be inserted following a geometric distribution of mean $1/(1-p_I)$ on the positive integers. Each base is drawn with equally likely probability $1/4$ from the set $\mathcal{A}$ and the index i is incremented accordingly.

   (b) A true copy occurs with probability $1 - p_I$. There is no effect on the fragment sequence.

The steps 1 and 2 are repeated independently for successive bases in the clone sequence and k is incremented until the fragment sequences reaches a

given length n_j. The entire process is repeated independently m times to produce the entire collection of fragment sequences. A small example is shown in figure 1.

3 The Assembled Fragments Matrix

We assume for purposes of this presentation that the collection of fragment sequences has been assembled into a matrix $\mathbf{X}$ with elements x_{ij}. Each row of this matrix contains the complete sequence of a fragment f_j, $j = 1, \ldots, m$. In addition, null characters "ϕ" may be inserted beyond the ends of a fragment sequence and gap characters "–" may be inserted internally. The total number of bases, gaps and null characters in each row of $\mathbf{X}$ is $n^\star$. Thus $\mathbf{X}$ is a matrix with m rows and $n^\star$ columns. The insertion of gaps and null characters defines a column–by–column relationship among the bases in each fragment sequence. This relationship is assumed to reflect the common origin of all fragment bases (or gaps) in a column from the same base (or bases) in the clone sequence. The columns will be denoted by $\mathbf{x}_i$. Algorithms for the assembly of fragment sequences have been described by several authors (Churchill et al., 1991; Kececioglu and Myers, 1990).

The gap characters inserted into a fragment sequence are of two types:

1. those generated by a deletion in the copying of a base,

2. those required to fill the spaces left by insertions in other fragments.

In general it will not be known to which class a gap character belongs. It is important to note that if all fragments spanning a base s_i produced deletions at that point, there would exist no corresponding column in the fragments matrix. Conversly, there may be columns in the fragments matrix which are generated by insertion errors and thus have no corresponding base in the clone sequence. For these reasons, the length of the clone sequence n and the width of the fragments matrix n^* are not neccessarily the same.

Just as for the clone sequence, we imagine the existence of links between the columns of the fragments matrix. We will refer to these as d–links. For each column $\mathbf{x}_i$ its associated d–link is immediately to the right. Each of the d–links corresponds either to an s–link or to one or more bases (plus their s–links) in the clone sequence. The latter situation arises when all of the fragments which cover a particular base(s) in the clone sequence have incurred a deletion of that base. In the case of insertion errors, one or more d–links (and their associated columns) may correspond to the same s–link.

The depth of a column $\mathbf{x}_i$ in the fragments matrix is defined to be the

number of non–null characters which it contains,

$$d_i = \sum_{j=1}^{m} \mathbf{1}(x_{ij} \neq \phi). \tag{1}$$

The depth of the d–link associated with column $\mathbf{x}_i$ is the minimum of d_i and d_{i+1}. The depth of the d–link associated with column x_{n^*} is zero.

4 Restoration of the DNA Sequence

In this section, we assume that the error rate parameters are known and describe an algorithm which generates samples from the distribution $\Pr(\mathbf{s} \mid \mathbf{X}, \theta)$.

Restoration Space We let $s_i^\star$ denote the base(s) associated with the column x_i and its d–link. Because the number of DNA bases associated with column x_i may be $0, 1, 2, \ldots$ we require the restoration to exist on a space with more general structure than $\mathcal{S}$. We let $\mathcal{A}^\star = \bigcup_{k=0}^{\infty} \mathcal{A}^k$, where $\mathcal{A}^0 = \{-\}$ is the gap character and $\mathcal{A}^k$ are the sets of k–tuples on $\mathcal{A}$ as before. When a column is generated by one or more insertion errors, the restored sequence should be $s_i^\star \in \mathcal{A}^0$. A column x_i generated by true copies and/or substitutions should be restored to $s_i^\star \in \mathcal{A}^1$ and the higher order sets $\mathcal{A}^k$, $k \geq 2$ occur when multiple deletion events are "hidden" in the d–link. The entire restored sequence $\mathbf{s}^\star$ exits on a space $\mathcal{R}$ which is the $n^\star$ Cartesian product of $\mathcal{A}^\star$. We will refer to

$\mathcal{R}$ as the restoration space. There is an obvious many–to–one mapping from $\mathcal{R}$ onto $\mathcal{S}$ which operates by removing gap characters and reindexing.

Prior Distribution We assume a prior distribution for $\mathbf{s}^\star$ which is independent across each component. Thus,

$$\Pr(\mathbf{s}^\star) = \prod_{i=1}^{n^\star} \Pr(s_i^\star). \tag{2}$$

The prior distribution is defined in two stages. First, given that $s_i^\star$ belongs to a particular size class, the distribution should be equally likely,

$$\Pr\left(s_i^\star = b \mid s_i^\star \in \mathcal{A}^k\right) = \frac{1}{4^k} \tag{3}$$

for all $b \in \mathcal{A}^k$ and $k = 0, 1, 2, \ldots$. The assignment of prior probabilities to each size class is somewhat more problematic. We will let $\gamma_0 = \Pr(s_i^\star \in \mathcal{A}^0)$ and distribute the remaining probability mass over the size classes $k = 1, 2, \ldots$ according to a geometric distribution with parameter γ_1. Thus, given $s_i \notin \mathcal{A}^0$

$$\Pr\left(s_i^\star \in \mathcal{A}^k\right) = (1 - \gamma_1)\gamma_1^{k-1} \tag{4}$$

for $k \geq 1$. Using this prior, the expected length of $\mathbf{s}^\star$ (when mapped onto $\mathcal{S}$) is

$$\mathrm{E}(n) = n^\star \frac{1 - \gamma_0}{1 - \gamma_1}. \tag{5}$$

Particular choices for the values of γ_0 and γ_1 will be motivated in the example.

Posterior Distribution Given the copying model of section 2 plus the assumption that the assembly of fragments is correct, we have

1. Mutual independence of the s_i given $\mathbf{X}$.

2. Independence of x_i and s_j for $i \neq j$.

Thus the posterior distribution can be factored

$$\Pr(\mathbf{s}^\star \mid \mathbf{X}, \theta) = \prod_{i=1}^{n^\star} \Pr(s_i \mid \mathbf{x}_i, \theta). \tag{6}$$

and we can restore s_i independently for each column.

The conditional distribution required to generate s_i can be computed using Bayes' rule

$$\Pr(s_i^\star \mid \mathbf{x}_i, \theta) \propto \Pr(s_i^\star \mid \theta) \prod_{j=1}^{m} \Pr(x_{ij} \mid s_i^\star, \theta). \tag{7}$$

The conditional probabilities of fragment elements given the clone sequence are

$$\Pr(x_{ij} \mid s_i^\star = -) = \begin{cases} \frac{1}{4} p_I, & x_{ij} \in \mathcal{A} \\ 1 - p_I, & x_{ij} = - \end{cases} \tag{8}$$

$$\Pr\left(x_{ij} \mid s_i^\star = \mathbf{b}, \mathbf{b} \in \mathcal{A}^k, k \geq 1\right) = \begin{cases} 1 - p_S - p_D, & x_{ij} = b_1 \\ \frac{1}{3} p_S, & x_{ij} \in \mathcal{A}, x_{ij} \neq b_1 \\ p_D, & x_{ij} = - \end{cases} \tag{9}$$

5 Estimation of the Error Rate Parameters

In this section, we assume that the clone sequence **s** and its correspondence with the fragments matrix are known and consider the problem of estimating the error rate parameters.

Prior Distribution The model as described in section 1 is defined in terms of the parameters $\theta = (p_D, p_S, p_I)$. We will assume the following prior distribution for θ:

$$\pi(p_D, p_S, p_I) = \pi_1(p_D, p_S) \cdot \pi_2(p_I) \tag{10}$$

where π_1 is Dirichlet with parameters $\alpha_1, \alpha_2, \alpha_3$ and π_2 is Dirichlet with parameters α_4, α_5. The choice of prior parameters may be based on previous DNA sequencing experiments or on subjective considerations.

Posterior Distribution The posterior distribution will again be a product of Dirichlets with parameters

$$\alpha_i^\star = \alpha_i + t_i \tag{11}$$

where t_i are the sufficient statistics

$$t_1 = \sum \mathbf{1}(x_{ij} = -, s_i \in \mathcal{A}) \tag{12}$$

$$t_2 = \sum \mathbf{1}(x_{ij} \in \mathcal{A}, s_i \in \mathcal{A}, x_{ij} \neq s_i) \tag{13}$$

$$t_3 = \sum \mathbf{1}(x_{ij} \in \mathcal{A}, s_i \in \mathcal{A}, x_{ij} = s_i) \quad (14)$$

$$t_4 = \sum \mathbf{1}(x_{ij} \in \mathcal{A}, s_i = -) \quad (15)$$

$$t_5 = \sum \mathbf{1}(x_{ij} = -, s_i = -). \quad (16)$$

The summations run over $i = 1, \ldots, n^\star$ and $j = 1, \ldots, m$. Note that t_1 counts d–links which correspond to a base in the clone sequence and t_5 counts d–links which are true copies of s–links.

6 A Resampling Algorithm

We have described, in sections 4 and 5, the conditional distributions $\Pr(\mathbf{s} \mid \mathbf{X}, \theta)$ and $\Pr(\theta \mid \mathbf{X}, \mathbf{s})$. We now describe an iterative resampling scheme which will generate observations from the marginal distributions $\Pr(\mathbf{s} \mid \mathbf{X})$ and $\Pr(\theta \mid \mathbf{X})$.

The Gibbs sampling algorithm is used to generate a sequence $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(k)}, \ldots$ with the property that $\theta^{(k)}$ is approximately a sample from $\Pr(\theta \mid \mathbf{X})$. Similarly, a sequence $\mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \ldots, \mathbf{s}^{(k)}, \ldots$ is also generated, where $\mathbf{s}^{(k)}$ is approximately a sample from $\Pr(\mathbf{s} \mid \mathbf{X})$. The approximation improves as k increases, becoming exact as $k \to \infty$. For a simple introduction to the Gibbs sampler see Casella and George (1991). For theoretical properties and examples see Gelfand and Smith (1990) and Gelfand et al. (1991).

At the j^{th} step, the Gibbs sampler generates $\theta^{(j)}$ and $\mathbf{s}^{(j)}$ according to

$$\begin{aligned} \theta^{(j)} &\sim \Pr\left(\theta \mid \mathbf{X}, \mathbf{s}^{(j-1)}\right) \\ \mathbf{s}^{(j)} &\sim \Pr\left(\mathbf{s} \mid \mathbf{X}, \theta^{(j)}\right). \end{aligned} \tag{17}$$

This iteration scheme generates two overlapping Markov chains whose stationary distributions are $\Pr(\theta \mid \mathbf{X})$ and $\Pr(\mathbf{s} \mid \mathbf{X})$, respectively. To implement the algorithm, we perform the iteration in 17 k times, obtaining our first values $\theta^{(k)}$ and $\mathbf{s}^{(k)}$, which we denote by $\theta_1^{(k)}$ and $\mathbf{s}_1^{(k)}$. This entire process is repeated N times, resulting in the values

$$\begin{aligned} &\theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_N^{(k)} \\ &\mathbf{s}_1^{(k)}, \mathbf{s}_2^{(k)}, \ldots, \mathbf{s}_N^{(k)}. \end{aligned} \tag{18}$$

For large k we treat the values in (18) as samples from $\Pr(\theta \mid \mathbf{X})$ and $\Pr(\mathbf{s} \mid \mathbf{X})$, respectively.

Using (18) to estimate probabilities is straightforward. For example, to estimate the probability that the clone sequence is a particular value $\mathbf{s}_0$, we calculate

$$\Pr(\mathbf{s} = \mathbf{s}_0 \mid \mathbf{X}) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\mathbf{s}_i^{(k)} = \mathbf{s}_0). \tag{19}$$

It has been noted (see for example Gelfand and Smith, 1991) that the estimate in (19) can be improved by applying the Rao–Blackwell theorem, which results

in the estimator

$$\Pr(\mathbf{s} = \mathbf{s}_0 \mid \mathbf{X}) \approx \frac{1}{N} \sum_{i=1}^{N} \Pr\left(\mathbf{s} = \mathbf{s}_0 \mid \mathbf{X}, \theta_i^{(k)}\right). \tag{20}$$

The calculation in (19) and (20) can similarly be done for probabilities about θ. The version in (20) is employed here.

Lastly, we again note two properties of these calculations. The expression in (20) becomes exact as k and $N \to \infty$. Thus by taking large enough values, we can attain any degree of accuracy in these calculations. Also, the calculation in (20) produces a probability that does depend on any estimated values of θ. Similarly, calculations about θ do not depend on any estimated values of s.

7 Summarizing the Sequence Sample

One result of the resampling algorithm is a collection of N sequences on the restoration space $\mathcal{R}$ which constitute a sample from the distribution $\Pr(\mathbf{s} \mid \mathbf{X})$.

A modal sequence is easily computed. For each column, $i = 1, \ldots, n^\star$, we choose the most frequently occuring element of $\mathcal{A}^\star$. Ties can be broken by randomization. The modal restoration can be mapped into sequence space and reported.

We would like to construct a set in the restoration space which contains

$100(1-\alpha)\%$ of the posterior probability. We consider cylinder sets, which are Cartesian products of subsets of $\mathcal{A}^\star$. Cylinder sets in restoration space can be mapped into cylinder sets in sequence space and are conveniently represented as a redundant DNA alphabet.

A highest posterior density (HPD) region can be found by choosing the smallest collection of sequences with total probability greater than $1-\alpha$. In practice, a proportion $(1-\alpha)$ of the sample is used to generate a cylinder set. The approximate HPD should be the smallest cylinder taken over such subsets of the sample. However, for any moderate to large sample, exhaustive search is not practical.

We suggest the following algorithm to find an approximate $100(1-\alpha)\%$ HPD region.

1. Compute a modal sequence $\mathbf{m}^\star$.

2. For every sequence in the sample, compute its distance from $\mathbf{m}^\star$ in Hamming measure on $\mathcal{R}$,

$$\delta_i = \sum_{j=1}^{n^\star} 1(m_j^\star \neq s_{ij}^\star),\ 1,\ldots,N. \tag{21}$$

3. Select the proportion $1-\alpha$ of sampled sequence which have the smallest distances δ_i.

4. Construct a cylinder set which contains all of the selected sequences. The cylinder should have components $r_i^\star = \bigcup_j s_{ij}^\star$, where the union is taken over all selected sequences.

For simplicity, we have reduced the cylinder set components to be $s_{ij}^\star$ when all the sampled sequences are identical at component i otherwise we note ambiguity of the restoration at column i.

This approximate HPD could be further refined by application of a stochastic annealing algorithm.

8 An Example

We illustrate the procedures using a set of sequenced fragments from *E. coli.*

9 Discussion

Our approach to estimation of DNA sequence accuracy depends on a number of assumptions which are not likely to be met in practice. We will briefly discuss some areas of concern and suggest potential generalizations which address these problems. Further discussion of the assumptions can be found in Churchill and Waterman (1991).

Error Rate Parameters We have described the model in terms of three error rate parameters and implicitly assumed (for example) that all possible substitutions of one base for another are equally likely. This simplification has allowed a clearer exposition of the methods and application of the algorithm to a small example. The generalization is straightforward. The parameters p_S and p_D can be expanded to a 4×5 stochastic matrix and the parameter p_I to 4–vector of probabilities. Such a model would have 19 free parameters and a product Dirichlet prior could be assumed.

A more serious shortcoming is that the error rates are assumed to be constant across the fragment sequences. It is well known that the error rates in fragment sequences increase as they are read out further on a gel. A model which incorporates position dependent error rates (e.g. $p_D(t)$, where t is a gel position index) is currently being developed, but presents a number of analytic difficulties.

The Assembled Fragments Matrix Our assumption that the assembled fragments matrix gives a true column–by–column correspondence is both crucial and troublesome. Fragment assembly is known to be a difficult and error prone process and there will inevitably be ambiguities in any assembly. One possible approach to overcome this problem is to consider a third level of sampling in the accuracy algorithm. Let F denote the unassembled fragments

and A the assembly. Note that F and A together determine X. The desired marginal distributions are $\Pr(S \mid F)$ and $\Pr(\theta \mid F)$. The resampling algorithm would procede in three steps

1. Generate a sequence from $\Pr(\mathbf{s} \mid F, A, \theta)$.

2. Generate an assembly from $\Pr(A \mid F, \mathbf{s}, \theta)$.

3. Generate parameters from $\Pr(\theta \mid F, A, \mathbf{s})$.

Steps 1 and 3 would procede exactly as described above. Step 2 requires the definition of a probability distribution over a space of sequence assemblies and a sampling algorithm for such a distribution. Implementation and theory of the Gibbs sampler are easily generalized to this case. Thus the difficulties in implementing this strategy lie in step 2.

Systematic Errors The assumptions of independent errors both within and between sequence fragments are troublesome and perhaps not easily addressed. Systematic errors in sequencing are known to occur and will require special attention in practical applications.

10 References

1. Berger, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis.* 2nd ed. Springer–Verlag

2. Casella and George Explaining the Gibbs sampler American Statistician 1992

3. Churchill, G.A., Burks, C., Eggert, M., Engle, M.L., Waterman, M.S. (1991) Fragment assembly methods for DNA sequencing. Manuscript.

4. Churchill, G.A. and Waterman, M.S. (1991). The accuracy of DNA sequences: estimating sequence quality. *Genomics* in press.

5. Kececioglu, J. and Myers, E. (1990). A robust automatic fragment assembly system. Manuscript.

6. Roberts, L. (1990). Large-scale sequencing trials begin. *Science*, **250**: 1336–1338.

7. States, D.J. and Botstein, D. (1991). Molecular sequence accuracy and the analysis of protein coding regions. Proc. Natl. Acad. Sci. USA **88**:5518–5522.

8. Waterman, M.S. (1990). Genomic sequence databases. *Genomics*, **6**: 700–701.